GAINING PRODUCTION RULES FOR A MARKOV BRAILLE TRANSLATION ALGORITHM

by

Hermann Kamp

When we contract a word in Braille the chief difficulty is whether the contraction is correct in the etymological sense of the word. Let me give an example: suppose, the ending -ion is correctly shortened in words as religion, section, union, commission, etc., its usage in the word "lion", however, would not be permitted by the majority of (German) Braille-readers up to now. Especially the teachers of the blind raise the objection that the reader would loose the touch of natural language if those contractions were allowed. This may be doubted for the sample quoted above, but it is beyond discussion that a misused contraction is a grievous obstacle for the Braille-reader, especially, if in a compound word one part of the contraction belongs to the first and one part of it to the second word.

A semantical analysis could clear the problem whether a certain group of letters is a syllable or a part-word-contraction permitted; but at least for the near future automatical semantics are rather a project of scientific research than of practical production. Transformational grammar may work successfully on a number of selected examples but until now it is far from processing language in its full scope.

Therefore, we propose a different way to determine whether a sequence of letters is a contraction permitted in Braille. Instead of trying to find a semantic pattern which might be valid for all cases of abbreviation we chose to deal separately with each syllable or part-word-contraction of German Braille.

For this purpose, we stored about 100,000 words, which were partly taken from critical comments on politics, arts, sport and so on, in our computer. Moreover, we could use a dictionary of about 130,000 words which Dr. Hubner from the IBM Company of Germany kindly made available to us. This combined material was analysed by a program to get an inventory of all words in which a certain sequence of letters occurred. Now it was checked in which cases this group of letters had not to be contracted in Braille, or - in other words - in which cases the given character string required a production rule of its own.

The next step was to examine the preceding and/or following letters, and to find a rule which describes the environment in which a certain sequence of letters must be translated different from the normal translation rule. Thus, it may be possible to reduce the semantical problems for our purpose to a formal scan of the adjacent letters.

Let me give an example again: We produced an inventory of all words in which the letters - ion - occurred and found that in about 99 percent it was correct to abbreviate it, except in words like: dionysisch, Ion, Pionier, Spion, Champion, Lampion, Zion, Radionetz. I found it helpful to sort the words according to the preceding and following letter of such a group, so that all -ion with a preceding P, for instance, are printed together because in this way it becomes easy to verify that a rule is free from contradictions. Now the production rule for this example:

	DIONYS ⟶	(D)	(I)	(O)	(Y)	(S)	
ß	ION ⟶	(ß)	(I)	(O)	(N)		
	PION ⟶	(P)	(I)	(O)	(N)		
	ZION ⟶	(Z)	(I)	(O)	(N)		
	RADION ⟶	(R)	(A)	(D)	(I)	(O)	(N)
	ION ⟶	(ION)					

From comparative linguistics we learn that the way of word-formation is similar in many languages, especially if they stem from a common root; so the English-American word friend/ly has its counterpart in German freund/lich, Dutch frunt/lijk, and Scandinavian frynt/lig, and if it is possible to describe when -lich can be contracted in German Braille we hope that this may be also possible for other languages.

The production rule for this example is simple because -lich- can be contracted in German Braille anywhere except in the beginning of a word and in compounds if one part of -lich- belongs to the first and one part to the second word of the compound.

```
(KALI/CHEMIE)
  ⠘LICH        (L) (ICH)
 KALICH        (K) (A) (L) (I) (C) (H)
```

(several other rules are left, being particular to German only.)

It is no doubt that the interpretation of the inventory - as I gave it for -ion and -lich- requires some time and labor, but we think it worth trying it.

I'll not conceal two main problems:

1. In one or the other case the number of production rules necessary may be very large even if until now this didn't become true.

2. How to deal with names (person- or place-names), Christian names can be taken into account but family- or place-names, especially when taken from a foreign language are difficult to handle. A fair percentage may be solved, however, by having production rules for person- or place-names of chief public interest.

On the other hand we think that new words are formed generally according to already existing types of word-formation and therefore are known to the program. Moreover, the system is very flexible; a production rule can be changed just by taking out a card or adding one without a change of the translation algorithm.